Comparison between YOLO and SSD Mobile Net for Object Detection in a Surveillance Drone

Mohit Phadtare
School of mechanical engineering
Dr. Vishwanath Karad World Peace University
Pune , India

Varad Choudhari
School of mechanical engineering
Dr. Vishwanath Karad World Peace University
Pune , India

Rushal Pedram
School of mechanical engineering
Dr. Vishwanath Karad World Peace University
Pune , India

Sohan Vartak
School of mechanical engineering
Dr. Vishwanath Karad World Peace University
Pune , India

.

***Abstract*—Recent advances in robotics and computer vision are leading to new applications for camera-equipped drones. One such application is the detection of aerial objects for surveillance. However, despite recent advances in the relevant literature, object recognition remains a challenging task in computer vision. There are many object detection algorithms for surveillance. This paper discusses and compares various systematic approaches that analyze images and determine whether the captured image or video contains vehicles, people, animals, or something else based on our requirements. This object recognition is carried out on images, videos, and real-time monitoring with two widely used machine learning algorithms: YOLO and Mobile NetSSD. Each model detects the presence of vehicles and the number of people in the image, which is judged based on its accuracy and smooth video processing. The performance of the two algorithms is determined for the detection of the objects in the frame and the precision of the detection.**

Keywords—Image recognition, object detection, artificial intelligence, YOLO, SSD Mobile Net

I. Introduction

Car accidents and natural disasters have always been the main contributors to the number of deaths from unnatural causes. One of the main reasons is that aid cannot reach the accident or disaster site in time due to road closures caused by traffic, floods, or landslides. Our surveillance drone uses object recognition to identify people who are trapped at the scene of the accident or disaster. According to our requirements, we will also be able to recognize the vehicles and some other objects and count these objects and we can probably track them. Object recognition consists of two parts: location and classification. The detection pipeline begins by extracting the selective features (hair, HOG, convolution layer) and then a locator or classifier is used to classify the object. These locators and classifiers are generally found in an image based on the area hint approach or the technique of sliding windows across the image. Methods like deformable part models (DPM) are paradigmatic for the sliding window approach and methods like RCNN use the area hint approach to generate bounding boxes and then run a classifier on the explained bounding boxes. Next, post-processing is done to filter out duplicate bounding boxes. The type of pipeline used in these methods is difficult to optimize and very complex, since each component is trained separately in these systems. But systems like YOLO have reshaped object recognition as a unique one-step regression problem by merging into a single network. Therefore, in systems like YOLO, the algorithm performs calculations on an image to predict where they are and classify these objects. MobileNetV2 has also shown good accuracy on low-latency, low-power models. This article compared the YOLOv3, YOLOv5s, and Mobile NetSSD V2 systems to identify the most suitable algorithm for the mask recognition system.

II. OBJECT DETECTION

A. YOLO(You Only Look Once)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the Microsoft Word, YOLO has been dominating its area for a long term and there was a chief step forward in May 2020. Two up to date and higher variations of YOLO had been brought one after the alternative. One changed into the YOLOv4 evolved via way of means of the traditional authors Joseph Redmon and Alexey Bochkovskiy, the alternative being the freshly launched YOLOv5 via way of means of Glenn Jocher. Not being the traditional creator of the YOLO series, this new launch changed into obtained with a few controversies, however skipping beyond it, the v5 version has proven a massive overall performance boom from its predecessors. However, YOLOv5 possessed hundreds of benefits in engineering. The particularly liked alternate being the use of Python language in preference to C as in its preceding variations. That makes set up and integration with IoT gadgets loads easier. In addition, the PyTorch network is likewise large than the Darknet network, because of this that that PyTorch will get hold of extra contributions and has a exceptional increase capacity withinside the future. Along with the improvement of YOLO in 2016, many item detection algorithms with specific strategies have done terrific achievements as well.

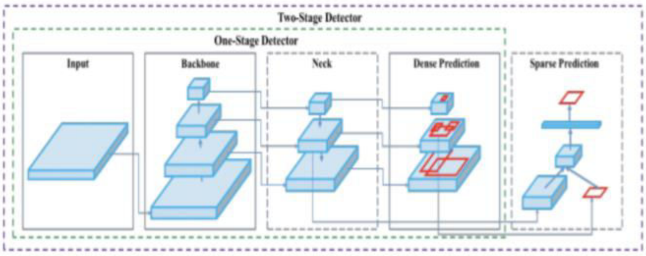

Figure 1

These improvements have formulated standards of architectural item detection: One-stage detector and Two-stage detector. Two concepts of architectural object detection.

The YOLOv5 network consists of three main parts.

a) Backbone - A CNN layer aggregate image features at different scales.
b) Neck – Set of layers to combine image features and pass them forward to prediction.
c) Head - Takes features from the neck and performs localization and classification.

The common point of all object detection architectures is that the input image features are compressed by a feature extractor (backbone) and then passed to the object detector (including the detection neck and detection head), as shown in Fig. 2 has the task of combining and mixing the features formed in the spine to prepare the next step in the sensing head (or head). The difference is that the head is responsible for detections along with the classification and location of each bounding box. The two-stage detector performs these 2 tasks separately and combines their results later (sparse detection), while the one-stage detector implements them at the same time (dense detection), also shown in Fig. 2. YOLO is a single stage detector, so you only see it once.

In the case of a single stage detector, the function of the head is to make dense predictions. The dense prediction is the final prediction, consisting of the prediction confidence value, the probability classes, and a vector containing the predicted bounding box coordinates (centre, height, width). YOLOv5 has an identical head to YOLOv3 for recognition

with anchor-based recognition steps and 3 levels of recognition granularity.

YOLOv5 comes in various versions, each having its own unique characteristic. These versions being:
1. yolov5-s - The small version
2. yolov5-m - The medium version
3. yolov5-l – The large version
4. yolov5-x - The extra-large version
The performance analysis of all these models as per Glenn Jocher is provided below. The extra-large version Below you will find the performance analysis of all these models according to Glenn Jocher.

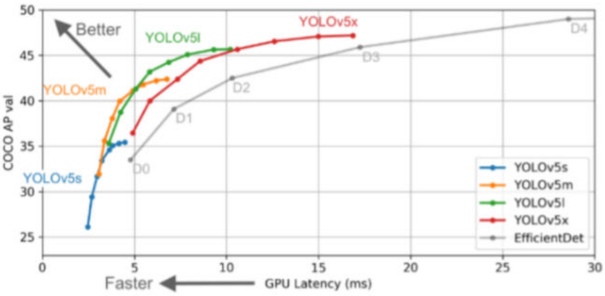

Figure 2

B. MobileNet-SSd V2*:*

The general trend that is being observed is that computer vision models are getting deeper and more complex to achieve greater precision. However, these advancements increase size and latency and cannot be used in systems with computational challenges. In such cases, MobileNet is useful. This is a model specially designed for mobile and embedded applications that require high speed. Its first version (MobileNetV1) had a deeply detachable fold that reduced the model size and cost of network complexity to an appropriate level to be useful for low-throughput applications. Accordingly, the second edition of the MobileNet family provides an inverted residual structure for significantly better modularity and this version is called MobileNetV2. This has helped eliminate non-linearities in tight layers, resulting in superior performance for previous applications. Around the time that the first version of MobileNet was introduced, Google released Single Shot Detector (SSD) for applications that rely heavily on speed and accuracy alike. As the name suggests, SSD essentially detected multiple objects in an image with a single shot. MobileNet is a model that offers decent speed, the only downside is its accuracy. SSD proved very useful for the model as it had the ability to improve accuracy while maintaining the speed of the models. The SSD algorithm was designed in such a way that it can be integrated into various networks such as YOLO, MobileNet, and the VGG architecture. Hence, MobileNet was incorporated into SSD for superior performance and was named Mobile Net-SSD. This integrated architecture is shown in Figure

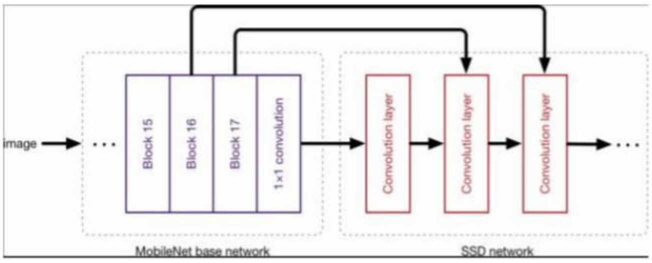

Figure 3 Integrated architecture of Mobile Net-SSD

III. Technology Stack:

A. Open CV:

OpenCV is one of the most popular computer vision libraries. If you want to start your journey in the field of computer vision, then a thorough understanding of the concepts of OpenCV is of paramount importance.
In this article, I will try to introduce the most basic and important concepts of OpenCV in an intuitive manner.

This article will cover the following topics:

1. Reading an image
2. Extracting the RGB values of a pixel
3. Extracting the Region of Interest (ROI)
4. Resizing the Image
5. Rotating the Image
6. Drawing a Rectangle
7. Displaying text

B. Darknet:

Darknet is an open source neural network framework. It is a fast and highly accurate (accuracy for custom trained model depends on training data, epochs, batch size and some other factors) framework for real time object detection (also can be used for images). The most important reason it is fast because it is written in C and CUDA.

C. Dataset:

Coco names: There is a file called coco.names that has the list of 80 object class that the model will be able to detect. The model has been trained only on these 80 object classes.

person	donut
bicycle	cake
car	chair
motorbike	sofa
aeroplane	pottedplant
bus	bed
train	diningtable
truck	toilet
boat	tvmonitor
traffic light	laptop
fire hydrant	mouse
stop sign	remote
parking meter	keyboard
bench	cell phone
bird	microwave
cat	oven
dog	toaster
horse	sink
sheep	refrigerator
cow	book
elephant	clock
bear	vase
zebra	scissors
giraffe	teddy bear
backpack	hair drier
umbrella	toothbrush
handbag	baseball glove
tie	skateboard
suitcase	surfboard
frisbee	tennis racket
skis	bottle
snowboard	wine glass
sports ball	cup
kite	fork
baseball bat	knife

sandwich	spoon
orange	bowl
broccoli	banana
carrot	apple
hot dog	pizza

IV. Comparison Results and Discussion

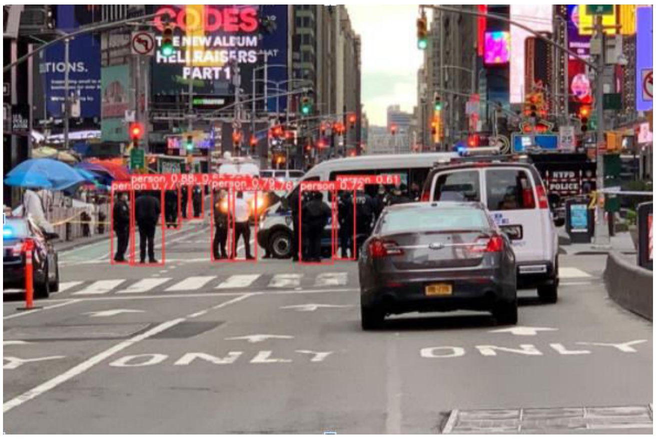

Figure 4.1 Person detection using YOLO v5

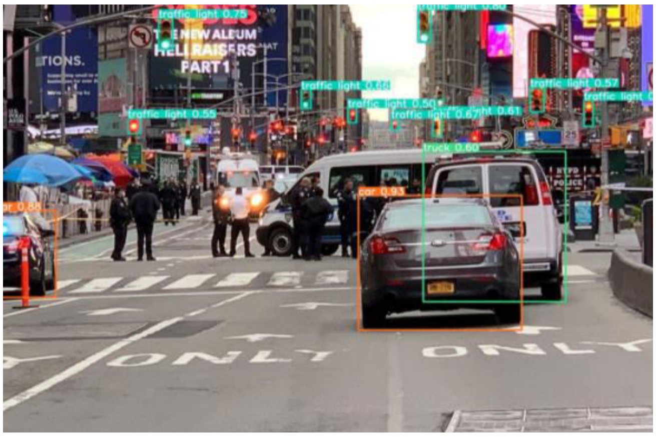

Figure 4.2 Vehicle detection using YOLO v5

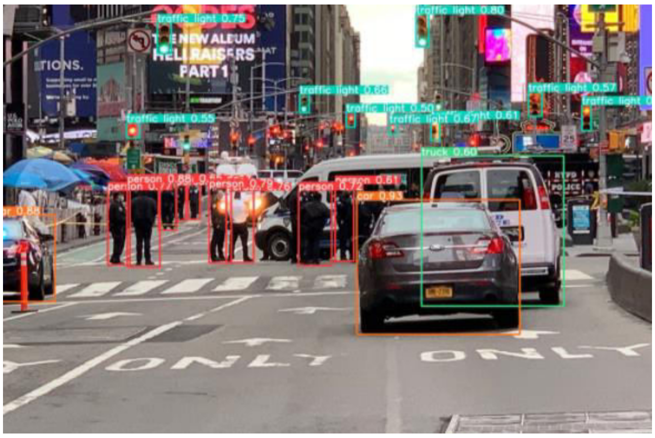

Figure 4.3 Person and Vehicle detection using YOLO v5

1. YOLO v5 (You only look once)

- For YOLO, recognition is a simple regression problem that takes an input image and learns the possibilities of the class with the bounding box coordinates. YOLO divides each image into an S x S grid and each grid predicts N bounding boxes and confidence.

- Confidence indicates the precision of the bounding box and whether the bounding box actually contains an object despite the defined class. YOLO even predicts the ranking score of each box for each class. You can merge the two classes to determine the probability that each class is present in a predicted box.

- In total, SxSxN boxes are forecast. On the other hand, most of these boxes have lower confidence values and if we set a gate threshold of, say, 30%, we can get rid of most of them.

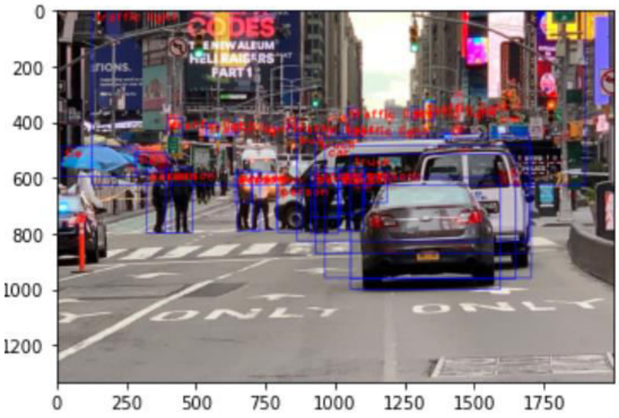

Figure 4.4 SSD Mobile net with same image

2. SSD Mobile-Net (Single Shot Detector)

- SSD strikes a better balance between speed and accuracy. SSD only performs a convolution network once on the input images and computes a feature map. We now run a small 3 × 3 large convolution kernel on this feature map to predict the bounding boxes and the probability of categorization.

- SSD also uses docking boxes in a variety of aspect ratios comparable to the Faster RCNN and learns scrolling to a degree that if it learns the box.

- To maintain scale, SSD predicts bounding boxes after multiple folding layers. Since each fold layer works at different scales, it can recognize objects at different scales.

- Taking into account the above points, it can be concluded that YOLOv5s is the most suitable model for real-time situations with optimal precision and fps values. It can be argued that Mobile Net-SSD V2 offers somewhat similar speed to YOLOv5s, but it simply lacks the precision department. For real-time purposes, speed is a crucial factor, but model accuracy is also essential for smooth operation. Then it can be said that one of the models can be selected based on the requirements of different applications.

V. CONCLUSIONS:

Therefore, two models were compared and each of them has its own unique properties. Each of the models was successful in the required mask recognition application, with YOLOv5s being the optimal model for real-time use due to its combination of speed and accuracy. The other is also a pretty decent model with different use cases. From this it can be concluded that machine vision applications can definitely be used in real time and that all these models are suitable to be converted into marketable products.

VI. REFERENCES

1. Deepa, R., et al. "Comparison of Yolo, SSD, Faster RCNN for Real Time Tennis Ball Tracking for Action Decision Networks." 2019 International Conference on Advances in Computing and Communication Engineering (ICACCE). IEEE, 2019.
2. Liu, Yifan, et al. "Research on the Use of YOLOv5 Object Detection Algorithm in Mask Wearing Recognition." World Scientific Research Journal 6.11 (2020): 276-284.
3. Jocher, G., Stoken, A., Borovec, J., Changyu, L., & Hogan, A. (2020). ultralytics/yolov5: v3. 0. Zenodo.
4. Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).
5. Chiu, Yu-Chen, et al. "Mobilenet-SSDv2: An improved object detection model for embedded systems." 2020 International Conference on System Science and Engineering (ICSSE). IEEE, 2020.
6. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.
7. R. Deepa, E. Tamilselvan, E. S. Abrar and S. Sampath, "Comparison of Yolo, SSD, Faster RCNN for Real Time Tennis Ball Tracking for Action Decision Networks," 2019 International Conference on Advances in Computing and Communication Engineering (ICACCE), 2019, pp. 1-4, doi: 10.1109/ICACCE46606.2019.9079965.
8. Hollemans, Matthijs. (2018, April 22). "MobileNet Version2." https://machinethink.net/blog/mobilenetv2/